

## PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-128396

(43)Date of publication of application : 16.05.1997

(51)Int.Cl.

G06F 17/28

(21)Application number : 07-287135

(71)Applicant : HITACHI LTD

(22)Date of filing : 06.11.1995

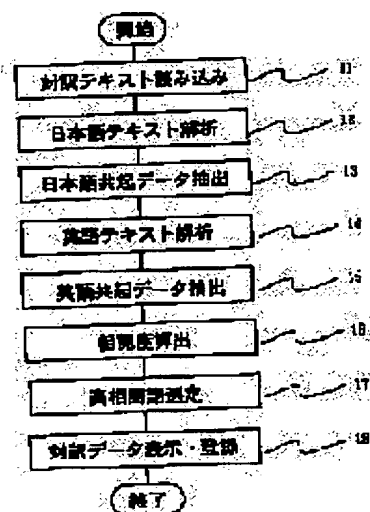
(72)Inventor : KAJI HIROYUKI  
ONO TOSHIKO

## (54) PREPARATION METHOD FOR BILINGUAL DICTIONARY

(57)Abstract:

PROBLEM TO BE SOLVED: To prepare a bilingual dictionary from texts where sentences are not made to correspond by extracting bilingual data of a word from the texts and automatically preparing the bilingual dictionary.

SOLUTION: The text of a first language and the text of a second language are read (11). The words appearing in the text of the first language are extracted (12) and the set of cooccurrence words is obtained for the respective words (13). The words appearing in the text of the second language are extracted (14) and the set of the cooccurrence words on the respective words are obtained (15). The correlation degree of the set of the cooccurrence words is calculated on the group of the words of the first language and the second language (16). The group of the words whose correlation degrees are largest is selected (17) and they are registered in the bilingual dictionary (18).



## LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-128396

(43) 公開日 平成9年(1997)5月16日

(51) Int.Cl.<sup>6</sup>

G 0 6 F 17/28

識別記号

庁内整理番号

F I

G 0 6 F 15/38

技術表示箇所

C

審査請求 未請求 請求項の数6 O L (全 17 頁)

(21) 出願番号 特願平7-287135

(22) 出願日 平成7年(1995)11月6日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 梶 博行

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 小野 敏子

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 弁理士 中村 純之助

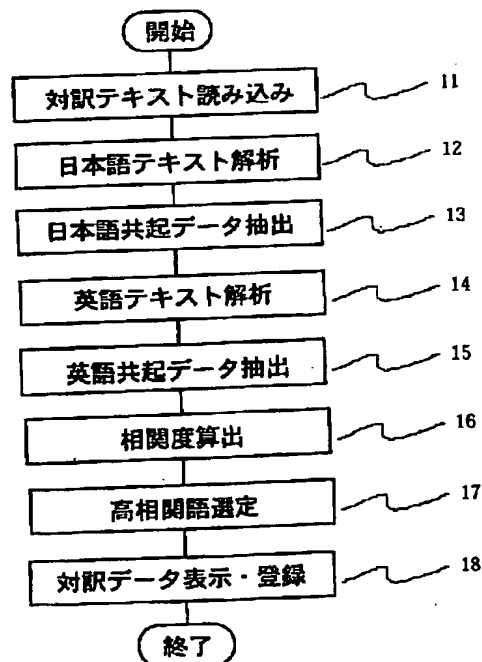
(54) 【発明の名称】 対訳辞書作成方法

(57) 【要約】

【課題】 対訳テキストから語の対訳データを抽出し、対訳辞書を自動生成する。これにより、文の対応づけがなされていない対訳テキストからの辞書作成を可能とする。

【解決手段】 第1言語のテキストと第2言語のテキストを読み込み(11)、第1言語のテキストに出現する語を抽出し(12)、各語について共起語の集合を求め(13)、第2言語のテキストに出現する語を抽出し(14)、各語について共起語の集合を求める(15)。第1言語の語と第2言語の語の組の各々について共起語集合の相関度を計算し(16)、互いに相関度が最大の語であるような語の組を選定し(17)、対訳辞書に登録する(18)。

図 2



## 【特許請求の範囲】

【請求項1】第1言語のテキストと第2言語のテキストを入力装置から読み込む対訳テキスト読み込みステップ、第1言語のテキスト中に出現する語を抽出する第1言語テキスト解析ステップ、第1言語のテキストに出現する語の各々についてテキスト中で共起する語の集合即ち第1の共起語集合を抽出する第1言語共起データ抽出ステップ、第2言語のテキスト中で出現する語を抽出する第2言語テキスト解析ステップ、第2言語テキストに出現する語の各々についてテキスト中で共起する語の集合即ち第2の共起語集合を抽出する第2言語共起データ抽出ステップ、第1言語の語の上記共起語集合と第2言語の語の上記共起語集合との相関度を計算する相関度算出ステップ、共起語集合の相関度に基づいて第1言語の語と第2言語の語の組を選定する高相関語選定ステップ、前記選定された語の組を対訳辞書に登録する対訳データ登録ステップから構成されることを特徴とする対訳辞書作成方法。

【請求項2】請求項1に記載の対訳辞書作成方法であって、相関度算出ステップは、対訳辞書に既登録の語の組を同一の要素とみなすことによって、第1言語の語の共起語集合と第2言語の語の共起語集合の相関度を計算することを特徴とする対訳辞書作成方法。

【請求項3】請求項1に記載の対訳辞書作成方法であって、高相関語選定ステップは、共起語集合の相関度が互いに最大の語であることを条件として、第1言語の語と第2言語の語の組を選定することを特徴とする対訳辞書作成方法。

【請求項4】請求項1に記載の対訳辞書作成方法であって、高相関語選定ステップは、対訳辞書の対訳データと語の出現頻度に基づく第2の相関度を算出し、選定する語に係わる第2の相関度より共起語集合の相関度が大きいことを選択の条件にすることを特徴とする対訳辞書作成方法。

【請求項5】請求項1に記載の対訳辞書作成方法であって、高相関語選定ステップは、共起語集合の相関度があらかじめ定めたしきい値以上であることを条件として、第1言語の語と第2言語の語の組を選定することを特徴とする対訳辞書作成方法。

【請求項6】請求項1に記載の対訳辞書作成方法であって、対訳データ登録ステップは、対訳辞書に登録する前に、語の組を表示装置に表示し、人間が登録を指示した語の組のみを対訳辞書に登録することを特徴とする対訳辞書作成方法。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】本発明は、機械翻訳システムなどで用いられる対訳辞書の作成方法に係わり、特に対訳テキストから語の対訳データを自動的に抽出する方法に関する。

## 【0002】

【従来の技術】機械翻訳システムにおいては対訳辞書が必須の構成要素であり、翻訳精度を向上するには、対訳辞書の語彙のカバー率を高めることが必要である。基本的な語彙の対訳辞書は、通常、機械翻訳システムのメーカーが提供するが、専門用語の対訳辞書はユーザが作成することが必要であり、辞書の作成コストが問題になる。このため、対訳テキストから語の対訳データを自動的に抽出することが望まれている。専門用語の対訳辞書は、機械翻訳システムだけでなく多言語情報検索システムなどでも必須の要素であり、自動作成のニーズは非常に高い。

【0003】対訳テキストから自動的に対訳辞書を作成する方法は、例えば、特開平7-28819号に開示されている。しかし、特開平7-28819号などの従来技術は、文の対応づけがなされた対訳テキストを利用するものであるという問題がある。対訳テキストは、多くの場合、テキスト全体で対訳になっているだけで、文単位での対応関係は付けられていないからである。従来技術によって対訳辞書を作成しようとする、対訳テキストにおける文の対応づけを行う前処理が必要になる。これを人手で行うのはコスト的に問題であり、文の対応づけを自動的に行う方法も研究されている。Computational Linguistics, Vol. 19, No. 1, pp. 75-102 (1993年3月)の論文"A Program for Aligning Sentences in Bilingual Corpora"はその例である。しかし、対訳テキストには、1つの文が2つの文に対応している部分も多いし、対応する文をもたない文が含まれることさえある。従って、文の対応づけを100%の精度で行うことは困難であり、コンピュータによる対応づけの結果を人間が確認・修正せざるを得ない。このように、文の対応づけコストを含めると、対訳辞書の作成コストの問題は解決されていないといえる。文の対応づけを前提としない対訳テキストからの対訳辞書作成方法としては、情報処理学会自然言語処理研究会報告No. 94-12 (1993年)の論文「対訳コーパスを用いた専門用語対訳辞書の作成」がある。しかし、これは、複数の単純語から構成される複合語を抽出し、単純語の対訳辞書を参照して構成要素間の対訳関係が確認できるような複合語の組を抽出する方法であり、単純語の対訳関係の抽出には適用できない。複合語であっても、構成要素の対応関係が明白で、かつ全ての対応関係が対訳辞書に含まれる場合しか抽出できないという問題がある。

## 【0004】

【発明が解決しようとする課題】本発明の目的は、上記従来技術の問題点を解決し、文の対応関係がつけられていない対訳テキストから、単純語と複合語の両方を対象として、対訳データを自動抽出する方法を提供することにある。

## 【0005】

3

【課題を解決するための手段】上記目的を達成するため、本発明の請求項 1 においては、対訳関係を有する第 1 言語のテキストと第 2 言語のテキストを入力装置から読み込む対訳テキスト読み込みステップ、第 1 言語テキストの形態素解析を行ってテキスト中に出現する語を抽出する第 1 言語テキスト解析ステップ、第 1 言語テキストの解析結果をもとに、テキスト中出现する語の各々について共起する語の集合即ち第 1 の共起語集合を抽出する第 1 言語共起データ抽出ステップ、第 2 言語テキストの形態素解析を行ってテキスト中出现する語を抽出する第 2 言語テキスト解析ステップ、第 2 言語テキストの解析結果をもとに、テキスト中出现する語の各々について共起する語の集合即ち第 2 の共起語集合を抽出する第 2 言語共起データ抽出ステップ、第 1 言語の語の第 1 の共起語集合と第 2 言語の語の第 2 の共起語集合との相関度を計算する相関度算出ステップ、共起語集合の相関度に基づいて第 1 言語の語と第 2 言語の語の組を選定する高相関語選定ステップ、前記選定された語の組を対訳辞書に登録する対訳データ表示・登録ステップから構成する。

【0006】請求項 2 においては、上記請求項 1 で述べた対訳辞書作成方法の相関度算出ステップにおいて、既に対訳辞書に登録されている語の組が存在する場合は、これを同一要素と見做すことによって、第 1 言語の語の共起語集合と第 2 言語の語の共起語集合の相関度計算を行う方法としている。

【0007】請求項 3 においては、上記請求項 1 で述べた対訳辞書作成方法の高相関語選定ステップにおいて、共起語集合の相関度が第 1 言語の語から見ても、第 2 言語の語から見ても最大となる語の組を対訳語として選定する方法としている。

【0008】請求項 4 においては、上記請求項 1 で述べた対訳辞書作成方法の高相関語選定ステップにおいて、対訳辞書の対訳データと語自身の対訳テキスト中での出現頻度に基づいて、上記共起語集合の相関度とは異なる第 2 の相関度を算出し、上記共起語集合の相関度の方が第 2 の相関度よりも大となっている語を選定条件としている。

【0009】請求項 5 においては、上記請求項 1 で述べた対訳辞書作成方法の高相関語選定ステップにおいて、共起語集合の相関度が予め定められた閾値以上となる第 1 言語の語及び第 2 言語の語の組を選定条件としている。

【0010】さらに請求項 6 においては上記請求項 1 で述べた対訳辞書作成方法の対訳データ登録ステップにおいて、データ処理により得られた対訳語の組を辞書に登録する前に一度表示装置で表示し、人間が確認した後に辞書への登録を行う方法としている。

【0011】

【発明の実施の形態】本発明の一実施例として、日英の

4

対訳テキストから語の対訳データを抽出する日英対訳辞書作成システムについて説明する。

【0012】日英対訳辞書作成システムのハードウェアは、図 1 に示すように処理装置 1、記憶装置 2、入力装置 3、表示装置 4 から構成される。処理装置 1 は対訳データを抽出する処理を実行する。記憶装置 2 は、日本語辞書 21、英語辞書 22、対訳辞書 23、日本語テキスト 24、英語テキスト 25 を格納するほか、対訳データ抽出処理の作業エリア 26 として用いられる。入力装置 3 は対訳テキストの入力に用いられ、表示装置 4 は抽出された対訳データの表示に用いられる。

【0013】処理装置 1 が実行する対訳データ抽出処理は、図 2 に示すように、対訳テキスト読み込みステップ 11、日本語テキスト解析ステップ 12、日本語共起データ抽出ステップ 13、英語テキスト解析ステップ 14、英語共起データ抽出ステップ 15、相関度算出ステップ 16、高相関語選定ステップ 17、対訳データ表示・登録ステップ 18 からなる。以下、各ステップについて説明する。

【0014】(1) 対訳テキスト読み込みステップ 11 対訳関係を有する日本語テキストと英語テキストを入力装置 3 から読み込み、記憶装置 2 の日本語テキスト 24 と英語テキスト 25 の格納エリアにそれぞれ格納する。

【0015】(2) 日本語テキスト解析ステップ 12 日本語テキスト 24 を読み出して文に分割し、さらに各文を語に分割する。併せて、複数の語から構成される複合語を抽出する。

【0016】テキストの文への分割は、テキストを構成する文字列を前方から 1 文字ずつチェックし、句点または改行記号が出現したら、それを文の末尾とみなすことによって行う。

【0017】文の語への分割は、日本語辞書 21 を参照して形態素解析することによって行う。形態素解析技術としては、例えば特開昭 61-40671 に開示されている技術を用いる。形態素解析の結果、文は語の列として表現されるが、本発明では、語の列のデータから助詞、助動詞などの機能語を除外し、名詞、動詞、形容詞、形容動詞などの内容語のみを残す処理を追加する。その理由は、機能語は言語間の対応関係が単純でなく、対訳テキストから抽出する対訳データを内容語の対訳関係に限定するのが適切であるからである。また、動詞など、活用する語はテキスト中にさまざまな変化形で出現するが、対訳辞書 23 に登録されている基本形（例えば、終止形）に置き換えて出力する。

【0018】複合語の抽出は、複合語を規定する品詞の並びを抽出することによって行う。例えば、連続する名詞の並びを複合名詞として抽出する。

【0019】日本語テキスト解析ステップ 12 によって得られる日本語テキスト解析結果 261a の例を図 3 (a) に示す。図 3 (a) において、「\」は語の区切

50

5

りを、「\」は文の区切りを、「\」はテキストの終了を示す。また、語のうしろの「(m, n)」は、当該語がテキスト中の第m字で始まり第n字で終わる語であることを表す。テキスト中の語の位置情報を付け加えた理由は、日本語共起データ抽出ステップ13において、語の重なりをチェックするためである。

【0020】(3) 日本語共起データ抽出ステップ13図3(a)における日本語テキスト解析結果261aをもとに、図4(a)における日本語出現語テーブル262および図5(a)における日本語共起頻度行列264 10を作成する。

【0021】日本語出現語テーブル262は、図4(a)に示すように、日本語テキストに出現した語2621とその出現頻度2622を示すテーブルである。なお、図4(a)の日本語出現語テーブルの内容は、図3(a)の日本語テキスト解析結果の内容に対応している。日本語共起頻度行列264は、図5(a)に示すように、日本語出現語テーブル262中の語に対応する行および列からなる行列であり、(i, j)要素は日本語出現語テーブル262中の第i語と第j語が同一文中に 20そろって出現した頻度を表す。なお、図5(a)の日本語共起頻度行列の内容は、図3(a)の日本語テキスト解析結果の内容に対応している。

【0022】日本語共起データ抽出ステップ13の処理を図6のフローチャートに沿って説明する。

【0023】最初に、日本語出現語テーブルのエントリ数(以後、簡単に「日本語語数」という)を表す変数を0に、日本語出現語テーブル262の語2621のフィールドを全て空白に、出現頻度2622のフィールドを 30全て0にする(1301)。また、図5(a)に示した日本語共起頻度行列264の全ての要素を0にする(1302)。さらに、日本語テキスト解析結果261a中の語を指すインデクスiに初期値1をセットする(1303)。

【0024】日本語テキスト解析結果261aから第i語を取り出してXにセットし(1304)、Xを引数にして日本語出現語テーブル検索/登録サブルーチンをコールする(1305)。日本語出現語テーブル検索/登録サブルーチンは、引数として与えられた語が日本語出現語テーブル262に登録されている場合は、そのエントリ番号をリターンし、引数の語が日本語出現語テーブル262中に未登録である場合は、テーブル末尾に登録した上でエントリ番号をリターンするサブルーチンである。日本語出現語テーブル検索/登録サブルーチンがリターンするXのエントリ番号をmにセットする(1306)。

【0025】次に、日本語テキスト解析結果261aの中の第i語のうしろの区切り記号を調べ(1307)。第i語が文の途中の語であれば、文中のそれ以降の語を取り出し、第i語との共起頻度を1増加する。そのた 50

6

め、まず、日本語テキスト解析結果261a中の語を指す第2のインデクスjに初期値として(i+1)をセットする(1308)。日本語テキスト解析結果261aから第j語を取り出してYにセットする(1309)。XとYがテキスト中で重なっていないかどうかチェックする(1310)。重なっていないければ、すなわち複合語とその構成要素のような関係でなければ、Yを引数にして日本語出現語テーブル検索/登録サブルーチンをコールし(1311)、日本語出現語テーブル検索/登録サブルーチンがリターンした後、日本語出現語テーブル262中のYのエントリ番号をnにセットする(1312)。m≠nであれば(1313)、日本語共起頻度行列の(m, n)要素および(n, m)要素をそれぞれ1ずつ増加する(1314)。m=nであれば(1313)、(m, n)要素を1だけ増加する(1315)。第i語と第j語の共起を処理したあと、日本語テキスト解析結果261a中の第j語のうしろの区切り記号を調べる(1316)。第j語が文の途中の語であれば、次の語との共起データを処理するため、jに1を加えて(1317)、1309に戻る。第j語が文末の語であれば、次の文の処理に進むため、iに1を加えて(1318)、1304に戻る。

【0026】なお、日本語テキスト解析結果261aの中の第i語のうしろの区切り記号をチェックする1307において、第i語がテキスト末の語であれば処理を終了する。また、第i語がテキスト末以外の文末の語であれば、次の文の処理に進むため、iに1を加えて(1318)、1304に戻る。

【0027】以上が日本語共起データ抽出ステップ13の処理である。次に、日本語共起データ抽出ステップの中でコールされる日本語出現語テーブル検索/登録サブルーチンの処理を図7のフローチャートに沿って説明する。まず、引数として与えられた語をWにセットし(13051)、日本語出現語テーブルの要素を指すインデクスkに初期値1をセットする(13052)。kが日本語語数以下であれば(13053)、Wを日本語出現語テーブル262の第k語と比較する(13054)。一致すれば、第k語の出現頻度2622を1だけ増加し(13055)、引数の語のエントリ番号としてkをリターンする(13056)。Wが日本語出現語テーブル262の第k語と一致しなければ(13054)、次の語と比較するため、kに1を加え(13057)、13053に戻る。kが日本語語数を越える時は(13053)、Wが日本語出現語テーブル262に未登録であることを意味するので、日本語出現語テーブル262の第k語としてWを登録し(13058)、日本語語数をkに更新する(13059)。このあと、第k語の出現頻度2622を1だけ増加し(13055)、引数の語のエントリ番号としてkをリターンする(13056)。

【0028】(4) 英語テキスト解析ステップ14

英語テキスト25を読み出して文に分割し、さらに各文を語に分割する。併せて、複数の語から構成される複合語を抽出する。

【0029】テキストの文への分割は、テキストを構成する文字列を前方から1文字ずつチェックし、ピリオドまたは改行記号が出現したら、それを文の末尾とみなすことによって行う。なお、この方法では、「Mr.」のようにピリオドで終わる語が出現すると誤って分割される。そのような語のリストを用意し、リスト中の語に關して例外処理をすることにより、分割精度を向上させることが可能である。

【0030】文の語への分割は、図1における英語辞書22を参照して形態素解析することによって行う。形態素解析技術としては、例えば特開昭58-40684号の中に開示されている技術を用いる。形態素解析の結果、文は語の列として表現されるが、本発明では、語の列のデータから前置詞、冠詞、助動詞などの機能語を除外し、名詞、動詞、形容詞、副詞などの内容語のみを残す処理を追加する。また、語はテキスト中にさまざまな変化形で出現するが、図1における対訳辞書23に登録されている基本形に置き換えて出力する。

【0031】複合語の抽出は、複合語を規定する品詞の並びを抽出することによって行う。例えば、連続する名詞の並びや、形容詞と後接する名詞の並びを複合名詞として抽出する。

【0032】英語テキスト解析ステップ14によって得られる英語テキスト解析結果261bの例を図3(b)に示す。英語テキスト解析結果261bに含まれる「\」、「\\」、「\\\」、および「(m, n)」の意味は日本語テキスト解析結果261aにおいてと同じである。

【0033】(5) 英語共起データ抽出ステップ15 英語テキスト解析結果261bをもとに、英語出現語テーブル263および英語共起頻度行列265を作成する。

【0034】英語出現語テーブル263は、図4(b)に示すように、英語テキストに出現した語2631とその出現頻度2632を示すテーブルである。なお、図4(b)の英語出現語テーブルの内容は、図3(b)の英語テキスト解析結果の内容に対応している。英語共起頻度行列265は、図5(b)に示すように、英語出現語テーブル263中の語に対応する行および列からなる行列であり、(i, j)要素は英語出現語テーブル263中の第i語と第j語が同一文中にそろって出現した頻度を表す。なお、図5(b)の英語共起頻度行列の内容は、図3(b)の英語テキスト解析結果の内容に対応している。

【0035】英語共起データ抽出ステップ15の処理は、日本語共起データ抽出ステップ13と全く同様であるので、詳細な説明は省略する。

【0036】(6) 相関度算出ステップ16

図4(a)、(b)に示した日本語出現語テーブル262、英語出現語テーブル263、図5(a)、(b)に示した日本語共起頻度行列264、英語共起頻度行列265、および図1における対訳辞書23に基づいて、図9に示す日英相関行列266を作成する。

【0037】対訳辞書23は、図8に例示するように、日本語の語231と英語の語232の組からなるレコードを記憶しており、日本語の語231をキーとして検索することができる。また、日英相関行列266は、図9に示すように、図4(a)に示した日本語出現語テーブル262中の語に対応する行、図4(b)に示した英語出現語テーブル263中の語に対応する列からなる行列であり、(i, j)要素は、共起語集合に基づく、日本語出現語テーブル262中の第i語と英語出現語テーブル263中の第j語の相関度を表す。図9の日英相関行列の内容は、対訳辞書の内容が図8であるとの前提で、図5(a)の日本語共起頻度行列と図5(b)の英語共起頻度行列から計算された結果である。

【0038】日本語の語JWと英語の語EWの相関度  $Assoc(JW, EW)$  は次式で定義する。

【0039】

$$Assoc(JW, EW) = C / (A + B - C)$$

ここに、 $A = JW$ の共起語集合の要素数、 $B = EW$ の共起語集合の要素数、 $C = JW$ の共起語集合と $EW$ の共起語集合の積集合の要素数。

【0040】ただし、ここでの集合は通常の集合と異なり、同一の語を複数個含むことを許し、集合の要素数は各語の個数の総和である。また、積集合は、 $JW$ の共起語集合中の語と $EW$ の共起語集合中の語の組が対訳辞書23に既に登録されているとき、これらの語を同一の要素とみなし、積集合を構成する要素と考える。また、同一とみなされる語の個数が二つの集合の間で異なるとき、積集合は少ないほうの個数を含むことにする。

【0041】例えば、図5(a)の日本語共起頻度行列264によれば「分割する」の共起語集合は次のとおりである。ここで、「/」のあとの数字が、「/」の前の語の個数を表している。

【0042】{日本語/1、テキスト/1、日本語テキスト/1、読み込む/1、文/2、語/1}

また、図5(b)の英語共起頻度行列265によれば「text」の共起語集合は次のとおりである。

【0043】{Japanese/1, read/1, divide/1, sentence/1}

ここで、対訳辞書が図8に示す3つのレコードのみを含むとすれば、「分割する」の共起語集合と「text」の共起語集合の積集合は次のようになる。ここでは、日本語の語と英語の語を＝で結ぶことによって、同一とみなされた要素であることを示している。

【0044】

9

(読み込む=read/1, 文=sentence/1)

従って、上の定義式による「分割する」と「text」の相関度は次のようになる。

【0045】 $Assoc(分割する, text) = 2 / (7 + 4 - 2) = 2 / 9$

相関度算出ステップ16は、図10に示すように、日英対訳行列作成サブステップ16a、日英仮想共起頻度行列計算サブステップ16b、日英相関行列計算サブステップ16cの3つのサブステップにわけられる。

【0046】日英対訳行列作成サブステップ16aは、  
図4(a)(b)に示した日本語出現語テーブル26  
2、英語出現語テーブル263、および対訳辞書23から日英対訳行列267を作成する。日英対訳行列267は、図11に示すように、日本語出現語テーブル262中の語に対応する行、英語出現語テーブル263中の語に対応する列からなる行列である。(i, j)要素の値は、日本語出現語テーブル262中の第i語と英語出現語テーブル263中の第j語の組が対訳辞書23に含まれているとき1、対訳辞書23に含まれていないとき0である。なお、図11の日英対訳行列の内容は図8の対  
20 訳辞書の内容に対応している。

【0047】日英対訳行列作成サブステップ16aの処理を図12のフローチャートに沿って説明する。

【0048】最初に、日英対訳行列267の全要素の値を0にする(1601)。次に、日本語出現語テーブル262の要素を指すインデックスiに初期値1をセットし(1602)、iが日本語語数と一致するまで(1613)、iに順次1を加えながら(1614)、以下の処理を繰り返す。

【0049】日本語出現語テーブル262中の第i語  
(以後、簡単に「日本語の第i語」という)をキーとして対訳辞書23を検索する(1603)。一つ以上の訳語が得られた場合(1604)、訳語を指すインデックスrに初期値1をセットし(1605)、rが得られた訳語数と一致するまで(1611)、順次rに1を加えながら(1612)、次の処理を繰返し実行する。英語出現語テーブル263の要素を指すインデックスjに初期値1をセットし(1606)、jが英語出現語テーブルのエントリ数(以後、簡単に「英語語数」という)と一致するまで(1607)、jに1を加えながら(1609)、第r訳語と英語出現語テーブル中の第j語(以後、簡単に「英語の第j語」という)を比較する動作(1608)を繰返す。一致するjに到達すると、日英対訳行列267の(i, j)要素の値を1にする(1610)。  
30

【0050】日英仮想共起頻度行列計算サブステップ16bは、日本語共起頻度行列264(図5(a))と日英対訳行列267(図11)から日英仮想共起頻度行列268(図13)を計算する。日英仮想共起頻度行列268は、図13に示すように、日本語出現語テーブル2  
50

10

62中の語に対応する行、英語出現語テーブル263中の語に対応する列からなる行列であり、(i, j)要素は日本語出現語テーブル262中の第i語と英語出現語テーブル263中の第j語との仮想的な共起頻度を表す。「仮想的な」共起とは、日本語テキストにおいて二つの語JW1とJW2が共起する場合、JW2の英訳語がJW1と共起するとみなすことを意味する。なお、図13の日英仮想共起頻度行列の内容は、図5(a)の日本語共起頻度行列と図11の日英対訳行列から計算された内容である。

【0051】日英仮想共起頻度行列計算サブステップ16bの処理を図14のフローチャートに沿って説明する。

【0052】日本語出現語テーブル262の要素を指すインデックスiに初期値1をセットし(1621)、iが日本語語数と一致するまで(1630)順次iに1を加えながら(1631)以下の処理を行い、さらにその過程において英語出現語テーブル263の要素を指すインデックスjに初期値1をセットし(1622)、jが英語語数と一致するまで(1628)jに1を加える操作を実行しつつ(1629)以下の処理を繰り返す。

【0053】日英仮想共起頻度行列268の要素の値を計算するための変数Xに初期値0をセットし、日本語出現語テーブル262の要素を指す第2のインデックスkに初期値1をセットする(1623)。日本語共起頻度行列264の(i, k)要素と日英対訳行列267の

(k, j)要素の積をXに加算する(1624)。kが日本語語数より小さければ(1625)、kに1を加え(1626)、1624に戻る。kが日本語語数に等しいならば、その時点のXの値を日英仮想共起頻度行列268の(i, j)要素の値として出力する(1627)。

【0054】日英相関行列計算サブステップ16c(図10)は、日本語共起頻度行列264、日英仮想共起頻度行列268と英語共起頻度行列265から日英相関行列266を計算する。このステップの処理を図15のフローチャートに沿って説明する。

【0055】日本語出現語テーブル262の要素を指すインデックスiに初期値1をセットし(1641)、iが日本語語数と一致するまで(1655)、順次iに1を加えながら(1656)以下の処理を繰り返す。

【0056】日本語の第i語の共起語集合の要素数を累計する変数Aに初期値0をセットし、日本語出現語テーブル262の要素を指す第2のインデックスkに初期値1をセットする(1642)。日本語共起頻度行列264の(i, k)要素をAに加える(1643)。kが日本語語数より小さければ(1644)、kに1を加え(1645)、1643に戻る。kが日本語語数に等しければ、その時点のAの値が、日本語の第i語の共起語集合の要素数を表している。

11

【0057】英語出現語テーブル263の要素を指すインデックスjに初期値1をセットし(1646)、jが英語語数と一致するまで(1653)順次jに1を加えながら(1654)以下の処理を繰り返す。

【0058】英語の第j語の共起語集合の要素数を累計する変数B、および日本語の第i語の共起語集合と英語の第j語の共起語集合の積集合の要素数を累計する変数Cに初期値0をセットし、英語出現語テーブル263の要素を指す第2のインデックスkに初期値1をセットする(1647)。英語共起頻度行列265の(j, k)要素をBに加える(1648)。また、日英仮想共起頻度行列268の(i, k)要素と英語共起頻度行列265の(j, k)要素の最小値をCに加える(1649)。kが英語語数より小さければ(1650)、kに1を加え(1651)、1648に戻る。kが英語語数に等しければ、その時点のBの値が、英語の第j語の共起語集合の要素数を表し、その時点のCの値が、日本語の第i語の共起語集合と英語の第j語の共起語集合の積集合の要素数を表している。以上のようにして得られたA、B、Cの値から $C/(A+B-C)$ を計算し、日英相関行列266の(i, j)要素として出力する(1652)。

【0059】(7)高相関語選定ステップ17

日英相関行列266および日英対訳行列267に基づいて、対訳データ269を抽出する。抽出される対訳データ269は、図16に例示するように、日本語の語2691、英語の語2692と相関度2693の組である。図16の対訳データは、図9の日英相関行列と図11の日英対訳行列から得られた対訳データである。

【0060】高相関語選定ステップ17の処理を図17、図18のフローチャートに沿って説明する。なお、図17と図18はL1、L2、L3でそれぞれ接続されている。

【0061】日本語出現語テーブル262の要素を指すインデックスiに初期値1をセットし(1701)、iが日本語語数と一致するまで(1726)順次iに1を加えながら(1727)、以下の処理を繰り返す。

【0062】英語出現語テーブル263の要素を指すインデックスjに初期値1をセットし、日本語出現語テーブルの第i語(以後、簡単に「日本語の第i語」という)に係わる相関度の最大値を記憶する変数Aの初期値を-1にする(1702)。日本語の第i語と英語出現語テーブルの第j語(以後、簡単に「英語の第j語」という)との相関度をAと比較し(1703)、Aより大であればAをその値に更新し、その時点のjの値をjmにセーブし、日本語の第i語との相関度がAである英語の語数を記憶する変数Nの値を1にする(1704)。日本語の第i語と英語の第j語との相関度がAと等しければ、Nを更新する(1705)。jが英語語数より小であれば(1706)、英語の次の語について処理するた

12

め、jに1を加え(1707)、1703に戻る。

【0063】英語出現語テーブルの全ての語について上記の処理が終了したら、日本語の第i語との相関度が最大値Aである英語の語数Nが1かどうかチェックする(1708)。

Nが1でなければ、日本語の第i語に係わる対訳データは抽出できなかったと判断し、1726に飛ぶ。Nが1であれば、日本語の第i語と英語の第jm語の組が対訳の候補になるので、英語の第jm語に係わる相関度とAとの大小をチェックする。すなわち、日本語出現語テーブル262の要素を指す第2のインデックスkに初期値1をセットし(1709)、kが日本語語数と一致するまで(1711)、kに1を加えながら

(1712)、日本語の第k語と英語の第jm語の相関度をAと比較する(1710)。英語の第jm語との相関度がAより大きな日本語の語があれば、日本語の第i語と英語の第jm語の組は対訳でないとして判断し、1726に飛ぶ。英語の第jm語との相関度がAより大きな日本語の語がなければ、日本語の第i語と英語の第jm語の組は、どちらの語からみても最大の相関度であるので、対訳の候補として残す。

【0064】次に、日本語の第i語と英語の第jm語の相関度Aを、日本語の第i語に係わる直接相関度、および英語の第jm語に係わる直接相関度と比較する。ここで、直接相関度とは、共起語集合に基づく相関度と異なり、対訳辞書23に対訳として登録されているかどうかということと、語自身の対訳テキスト中での出現頻度に基づく相関度である。

【0065】英語出現語テーブル263の要素を指すインデックスjに初期値1をセットし(1713)、英語語数と一致するまで(1717)順次jに1を加えながら(1718)、日本語の第i語と英語の第j語の直接相関度BをAと比較する。すなわち、日英対訳行列267の(i, j)要素が1であれば(1714)、日本語の第i語の出現頻度2622と英語の第j語の出現頻度2632の最小値を最大値で除した値をBとし(1715)、AとBとの大小を比較する(1716)。日本語の第i語との直接相関度BがA以上である英語の語があれば、日本語の第i語と英語の第jm語の組は対訳でないとして判断し、1726に飛ぶ。

【0066】同様に、日本語出現語テーブル262の要素を指す第2のインデックスkに初期値1をセットし(1719)、kが日本語語数と一致するまで(1723)、kに1を加えながら(1724)、日本語の第k語と英語の第jm語の直接相関度BをAと比較する。すなわち、日英対訳行列267の(k, jm)要素が1であれば(1720)、日本語の第k語の出現頻度2622と英語の第jm語の出現頻度2632の最小値を最大値で除した値をBとし(1721)、AとBとの大小を比較する(1722)。英語の第jm語との直接相関度BがA以上である日本語の語があれば、日本語の第i語



と英語の第  $j$  語の組は対訳でないとは判断し、1726に飛ぶ。

【0067】日本語の第  $i$  語あるいは英語の第  $j$  語に関し、 $A$  より大きな直接相関度をもつ語がない場合には、日本語の第  $i$  語、英語の第  $j$  語、および相関度  $A$  の組を対訳データ269として出力する(1725)。

【0068】(8) 対訳データ表示・登録ステップ18高相関語選定ステップ17で選定された対訳データ269を表示装置4に表示する。ユーザは、表示された対訳データの各々について、対訳辞書23に登録するか否かを入力装置3から指示することができる。対訳データの登録が指示されると、当該対訳データの日本語の語2691と英語の語2692を組にして対訳辞書23に登録する。

【0069】以上の(1)から(8)のステップを実行することにより、対訳テキストから語の対訳データを抽出し、対訳辞書を充実させていくことができる。例えば、図8に示す3つの対訳データから成る対訳辞書を利用して、図3(a)と図3(b)に示す対訳テキストを処理することにより、図16に示す2つの対訳データが対訳辞書に追加される。

【0070】以上、説明したのは一実施例であり、各ステップに種々のバリエーションを考えることができる。

【0071】「共起する語」について、上記実施例では「同一文中に出現する語」としたが、大きさ  $n$  のウィンドウに含まれる語を採用してもよい。例えば、大きさ7のウィンドウの場合、ある語と共起する語とは、その語の前後それぞれ3語の範囲に出現する語である。また、構文的に関係のある語(修飾/被修飾の関係にある語)を共起する語とする方法も考えられる。

【0072】共起データ抽出ステップにおいて、上記実施例では、複合語とその構成語の関係は共起関係ではないので、複合語とその構成語の組は共起頻度としてはカウントしていない。しかし、複合語と構成語は、共起とは違った意味で関連がある。すなわち、構成要素の間の対訳関係は、複合語の対訳関係抽出の手掛かりになる。従って、複合語の共起語の中にその構成語を含めて相関度を計算する方法も考えられる。この場合、相関度は、共起語の対訳知識だけでなく構成語の対訳知識を反映したものになる。

【0073】共起語集合の相関度についても、上記実施例以外に種々の定義が可能である。例えば、共起特性をベクトルで表現し、ベクトル間の角度が小さいほど相関が高いとする方法が考えられる。ここで、日本語の語の共起特性は、各成分が日本語の語に対応するベクトルで表現され、一方、英語の語の共起特性は、各成分が英語の語に対応するベクトルで表現される。従って、語の対訳関係に基づいて成分を対応づけた上でベクトル間の角度を計算することにする。

【0074】高相関語選定ステップにおいて、上記実施

例では、相関度が最大の語の組であっても、より大きな直接相関度をもつ語が存在する場合は除外している。この処理において、直接相関度に適当な重みをつけることが考えられる。また、直接相関度との比較処理を完全に省略することも考えられる。さらに、簡単に、あらかじめ定めたしきい値以上の相関度をもつ全ての語の組を選定する方法も考えられる。

【0075】

【発明の効果】本発明によれば、対訳辞書に既登録の対訳データを利用して、対訳辞書に未登録の対訳データを対訳テキストから自動的に抽出し、対訳辞書を充実させていくことができる。従来技術と異なり、文の対応がつけられていない対訳テキストから対訳データを抽出できることが本発明の顕著な効果である。

【図面の簡単な説明】

【図1】日英対訳辞書作成システムのハードウェア構成図。

【図2】対訳データ抽出処理のフローチャート。

【図3】(a)は日本語テキスト解析結果の例を示す図、(b)は英語テキスト解析結果の例を示す図。

【図4】(a)は日本語出現語テーブルの例を示す図、(b)は英語出現語テーブルの例を示す図。

【図5】(a)は日本語共起頻度行列の例を示す図、(b)は英語共起頻度行列の例を示す図。

【図6】日本語共起データ抽出処理のフローチャート。

【図7】日本語出現語テーブル検索/登録サブルーチンのフローチャート。

【図8】対訳辞書の例を示す図。

【図9】日英相関行列の例を示す図。

【図10】相関度算出処理のフローチャート。

【図11】日英対訳行列の例を示す図。

【図12】日英対訳行列作成処理のフローチャート。

【図13】日英仮想共起頻度行列の例を示す図。

【図14】日英仮想共起頻度行列計算処理のフローチャート。

【図15】日英相関行列計算処理のフローチャート。

【図16】抽出された対訳データの例を示す図。

【図17】高相関語選定処理のフローチャート(その1)。

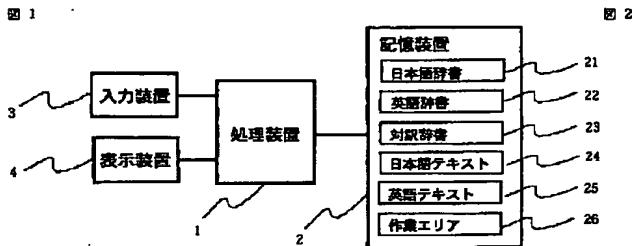
【図18】高相関語選定処理のフローチャート(その2)。

【符号の説明】

- |                   |                  |
|-------------------|------------------|
| 1 処理装置            | 2 記憶装置           |
| 3 入力装置            | 4 表示装置           |
| 11 対訳テキスト読み込みステップ | 12 日本語テキスト解析ステップ |
| 13 日本語共起データ抽出ステップ | 14 英語テキスト解析ステップ  |
| 15 英語共起データ抽出ステップ  | 16 相関度算出ステップ     |

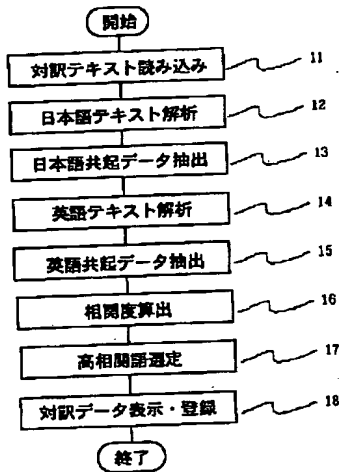
- 15
- 17 高相関語選定ステップ  
タ表示・登録ステップ
- 21 日本語辞書
- 23 対訳辞書  
キスト
- 25 英語テキスト  
ア
- 261a 日本語テキスト解析結果    261b 英語テキスト解析結果
- 18 対訳デー  
タ
- 22 英語辞書
- 24 日本語テ  
キスト
- 26 作業エリ  
ア

【図 1】

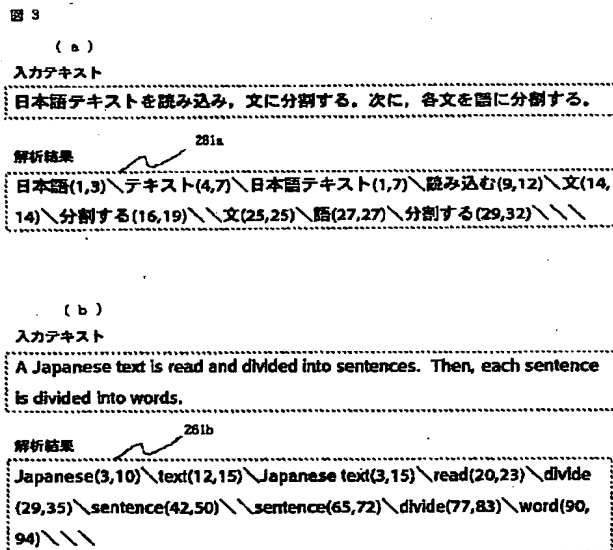


- 16
- 262 日本語出現語テーブル  
現語テーブル
- 264 日本語共起頻度行列  
起頻度行列
- 266 日英相関行列
- 268 日英仮想共起頻度行列  
れた対訳データ
- 263 英語出  
現語テーブル
- 265 英語共  
起頻度行列
- 267 日英対  
訳行列
- 269 抽出さ  
れた対訳データ

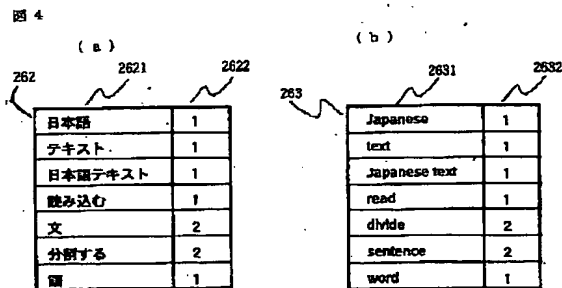
【図 2】



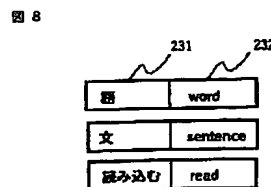
【図 3】



【図 4】



【図 8】



【図 5】

図 5

( a )

	日本語	テキスト	日本語 テキスト	読み込む	文	分割する	語
日本語	0	1	0	1	1	1	0
テキスト	1	0	0	1	1	1	0
日本語 テキスト	0	0	0	1	1	1	0
読み込む	1	1	1	0	1	1	0
文	1	1	1	1	0	2	1
分割する	1	1	1	1	2	0	1
語	0	0	0	0	1	1	0

264

【図 9】

図 9

Japanese text    Japanese read    divide    sentence    word

266

日本語	1/3	1/3	2/5	1/8	2/9	1/10	1/5
テキスト	1/3	1/3	2/5	1/8	2/9	1/10	1/5
日本語 テキスト	2/5	2/5	1/2	1/7	1/4	1/9	1/4
読み込む	1/8	1/8	1/7	1/9	1/11	0	1/6
文	1/10	1/10	1/9	0	1/6	1/6	0
分割する	2/9	2/9	1/4	1/11	2/5	1/6	1/8
語	1/5	1/5	1/4	1/6	1/8	0	1/3

266

( b )

265

	Japanese text	Japanese read text	divide	sentence	word		
Japanese	0	1	0	1	1	0	
text	1	0	0	1	1	0	
Japanese text	0	0	0	1	1	0	
read	1	1	1	0	1	0	
divide	1	1	1	1	0	2	1
sentence	1	1	1	1	2	0	1
word	0	0	0	0	1	1	0

265

【図 11】

図 11

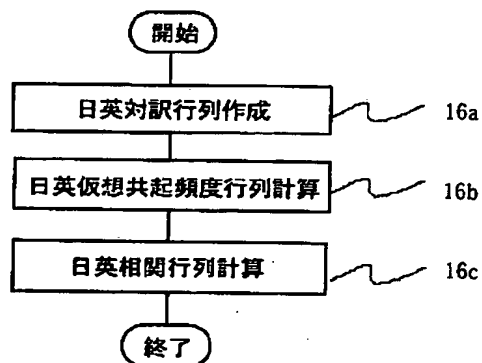
	Japanese text	Japanese read text	divide	sentence	word
日本語	0	0	0	0	0
テキスト	0	0	0	0	0
日本語 テキスト	0	0	0	0	0
読み込む	0	0	1	0	0
文	0	0	0	0	1
分割する	0	0	0	0	0
語	0	0	0	0	1

267

267

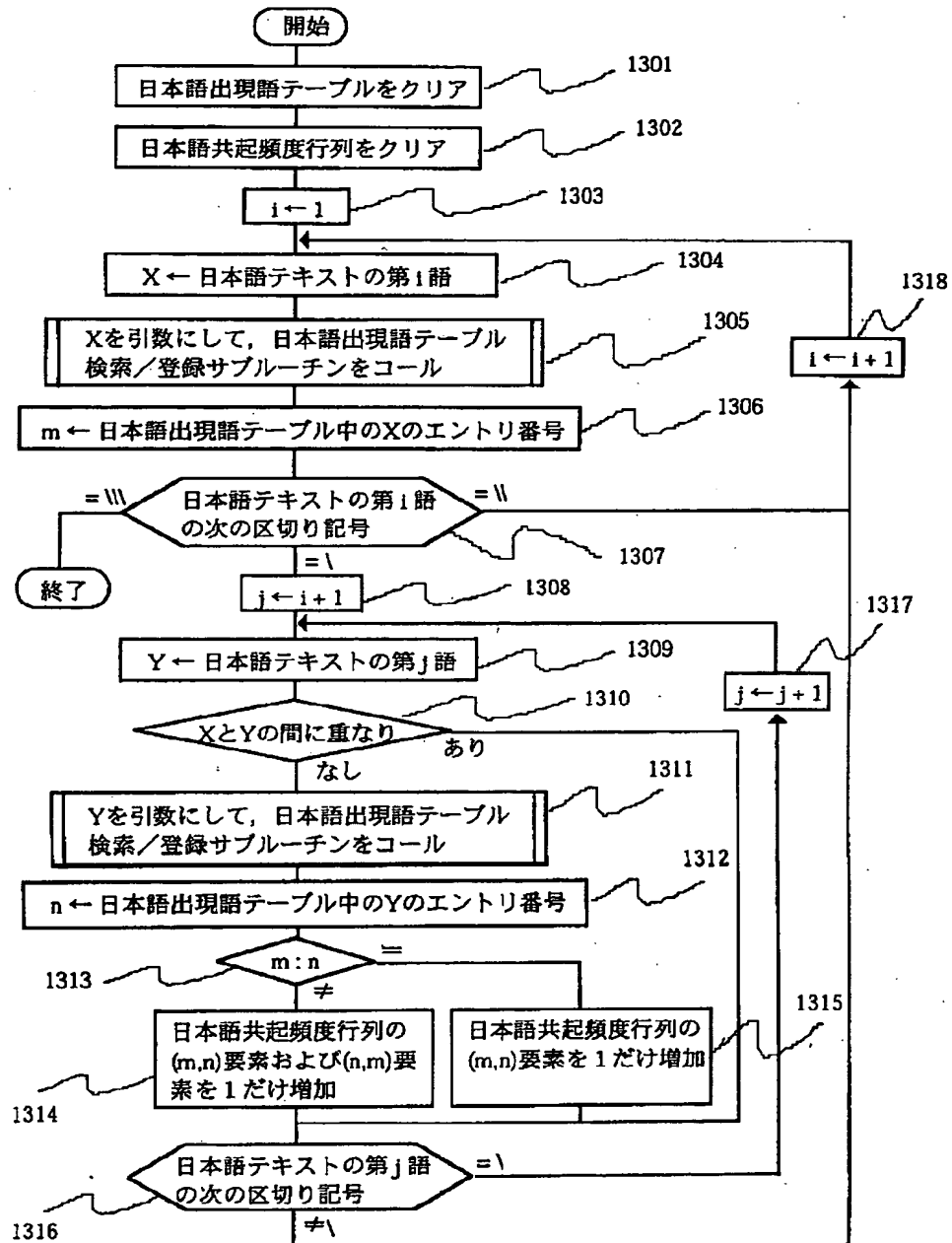
【図 10】

図 10



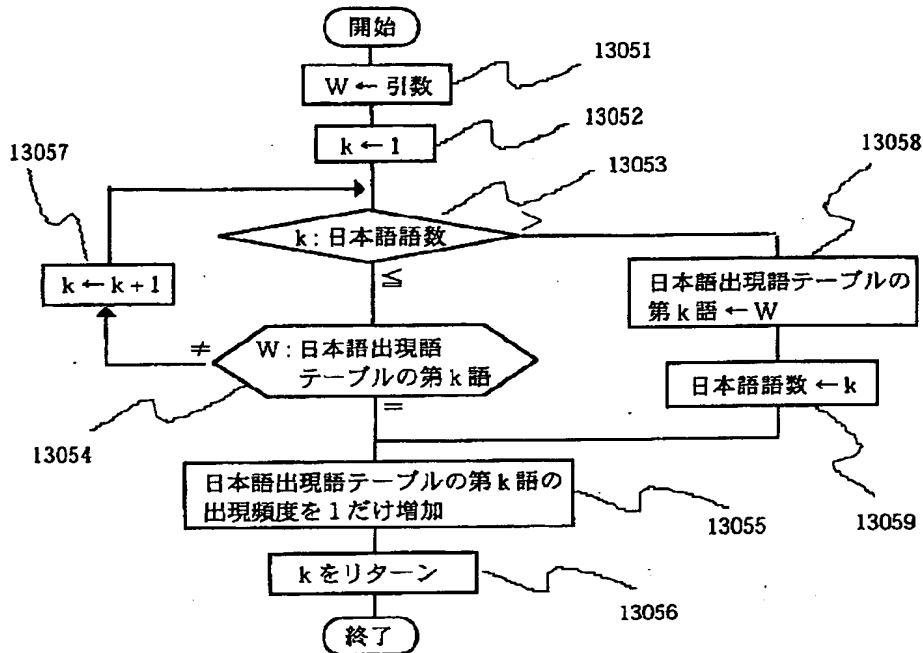
【図 6】

図 6



【図 7】

図 7



【図 13】

図 13

	Japanese text	Japanese read	divide	sentence	word	
日本語	0	0	0	1	0	1
テキスト	0	0	0	1	0	1
日本語	0	0	0	1	0	1
テキスト	0	0	0	0	0	1
読み込む	0	0	0	1	0	0
文	0	0	0	1	0	0
分割する	0	0	0	1	0	2
語	0	0	0	0	0	1

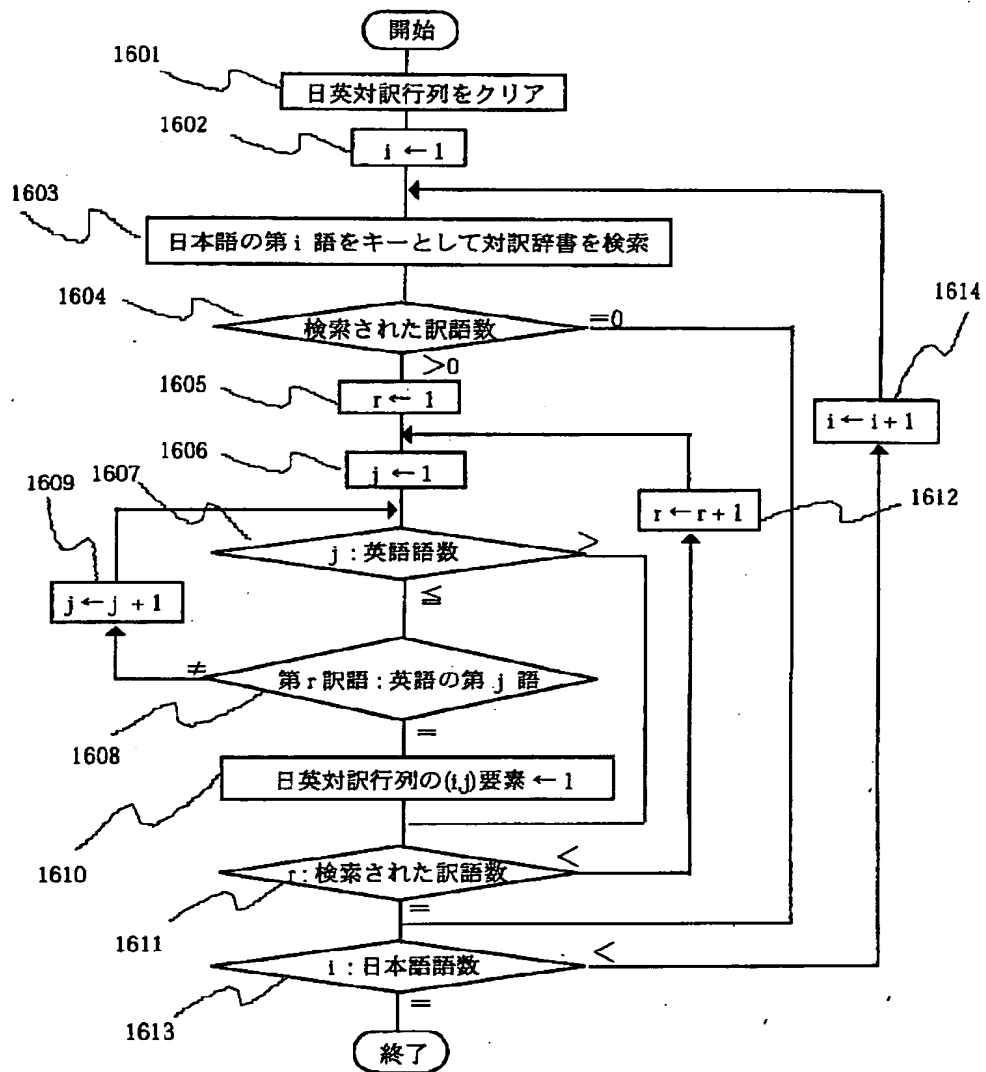
【図 16】

図 16

日本語テキスト	Japanese text	1/2
分割する	divide	2/5

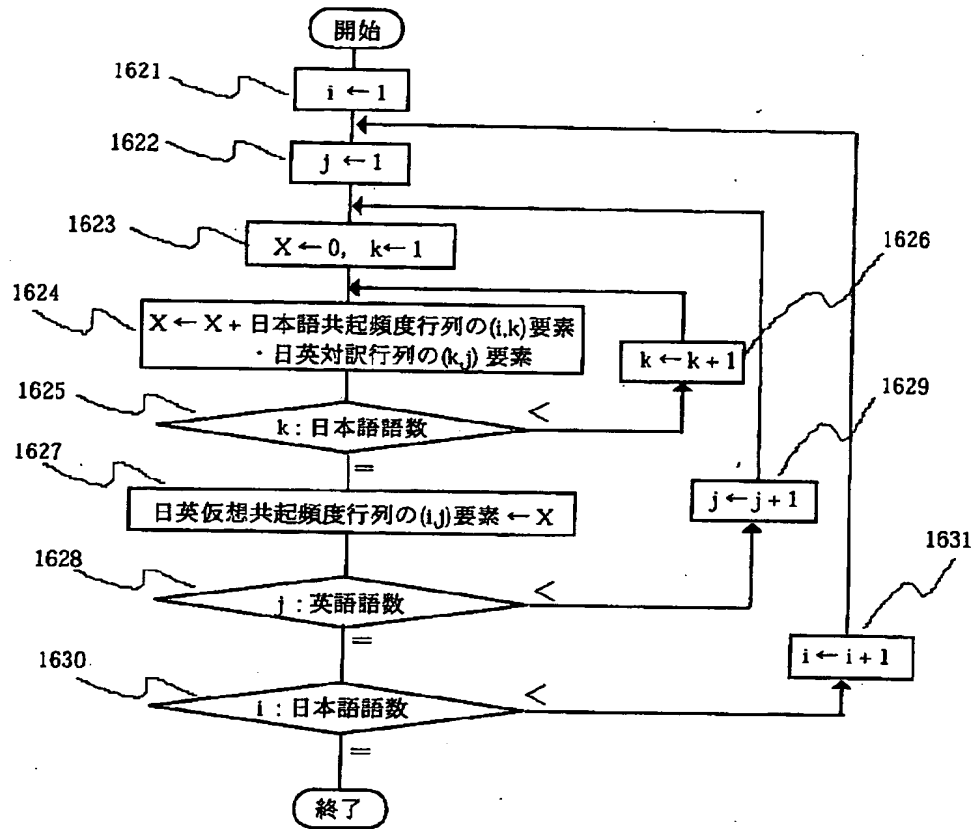
【図12】

図 1 2



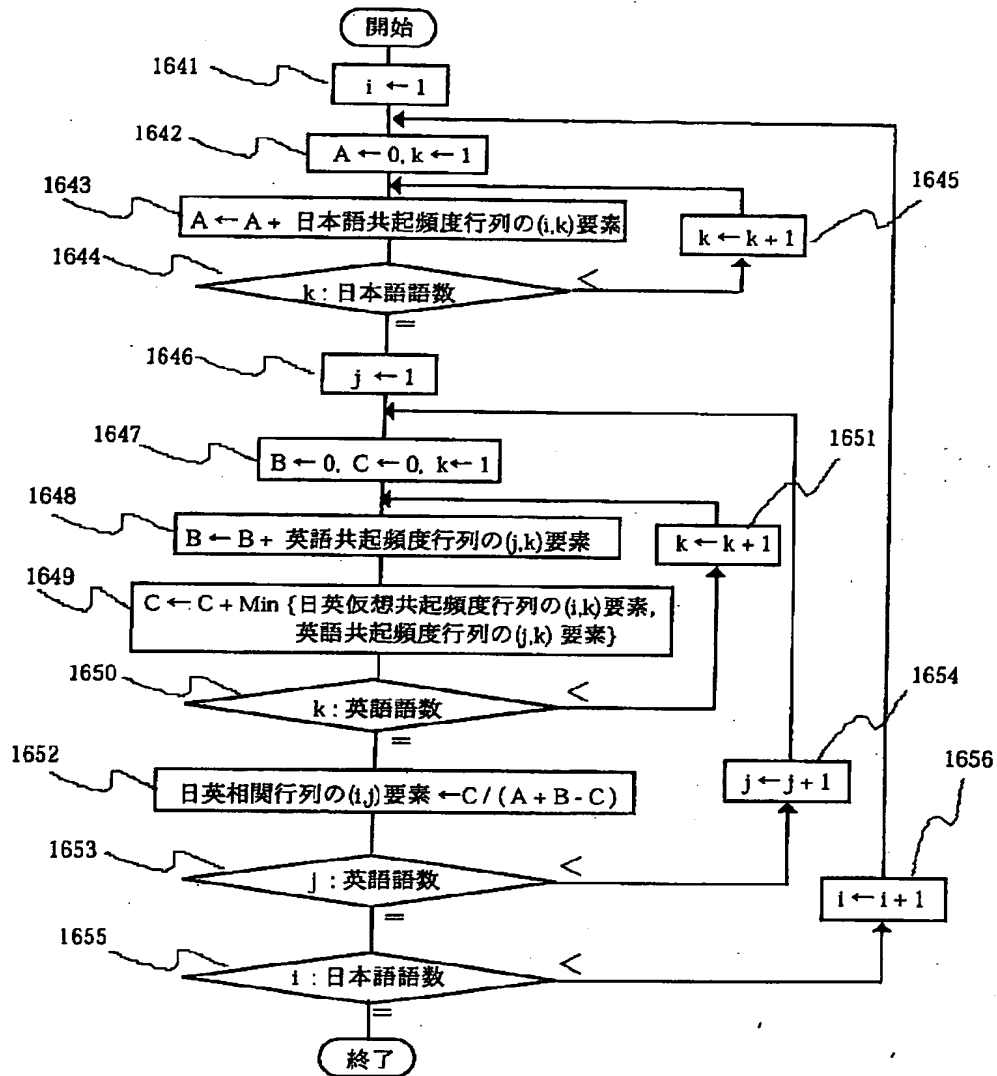
【図 1 4】

図 1 4



【図 1 5】

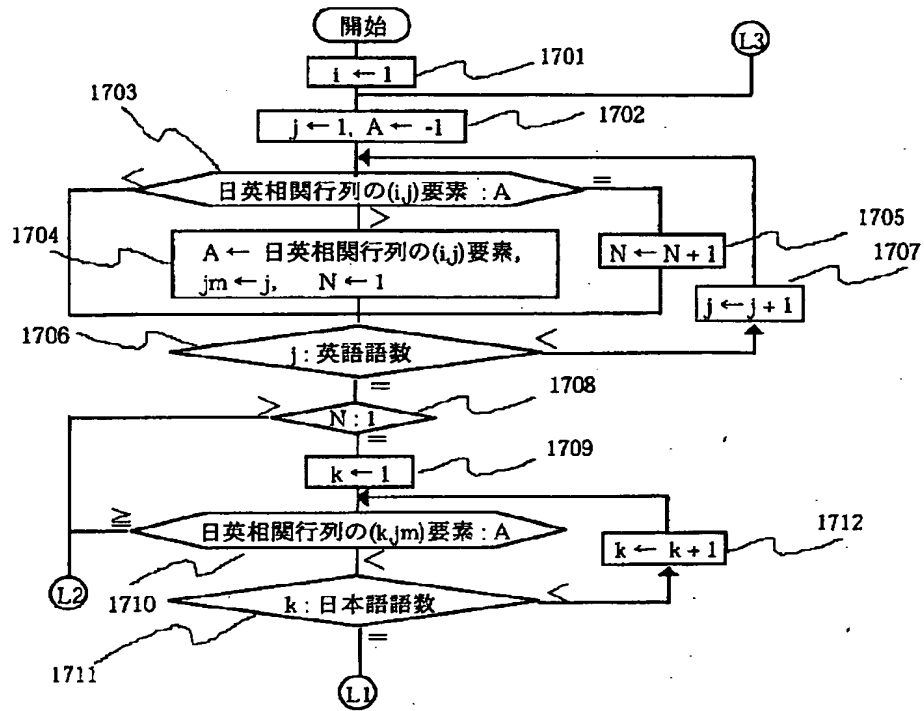
図 1 5





【図17】

図 1 7



【図 18】

図 18

